In [1]:
```
import nltk
import numpy as np
```

Open corpus, define "documents"

In [2]:
```
myfile = open("english.txt", "r")
```

In [3]:
```
corpus = myfile.read()
```

In [4]:
```
print(corpus)
```

Whereas recognition of the inherent dignity and of the equal and inalie
nable rights of all members of the human family is the foundation of fr
eedom justice and peace in the world
Whereas disregard and contempt for human rights have resulted in barbar
ous acts which have outraged the conscience of mankind and the advent o
f a world in which human beings shall enjoy freedom of speech and belie
f and freedom from fear and want has been proclaimed as the highest asp
iration of the common people
Whereas it is essential if man is not to be compelled to have recourse
as a last resort to rebellion against tyranny and oppression that human
rights should be protected by the rule of law
Whereas it is essential to promote the development of friendly relation
s between nations
Whereas the peoples of the United Nations have in the Charter reaffirme
d their faith in fundamental human rights in the dignity and worth of t
he human person and in the equal rights of men and women and have deter
mined to promote social progress and better standards of life in larger
freedom
Whereas Member States have pledged themselves to achieve in cooperation
with the United Nations the promotion of universal respect for and obse
rvance of human rights and fundamental freedoms
Whereas a common understanding of these rights and freedoms is of the g
reatest importance for the full realization of this pledge
Proclaims this Universal Declaration of Human Rights as a common standa
rd of achievement for all peoples and all nations to the end that every
individual and every organ of society keeping this Declaration constant
ly in mind shall strive by teaching and education to promote respect fo
r these rights and freedoms and by progressive measures national and in
ternational to secure their universal and effective recognition and obs
ervance both among the peoples of Member States themselves and among th
e peoples of territories under their jurisdiction
All human beings are born free and equal in dignity and rights
They are endowed with reason and conscience and should act towards one
another in a spirit of brotherhood
Everyone is entitled to all the rights and freedoms set forth in this D
eclaration without distinction of any kind such as race colour sex lang
uage religion political or other opinion national or social origin prop
erty birth or other status
Furthermore no distinction shall be made on the basis of the political
jurisdictional or international status of the country or territory to w
hich a person belongs whether it be independent trust non self governin
g or under any other limitation of sovereignty
Everyone has the right to life liberty and security of person
No one shall be held in slavery or servitude
slavery and the slave trade shall be prohibited in all their forms
No one shall be subjected to torture or to cruel inhuman or degrading t
reatment or punishment
Everyone has the right to recognition everywhere as a person before the
law
All are equal before the law and are entitled without any discriminatio
n to equal protection of the law
All are entitled to equal protection against any discrimination in viol
ation of this Declaration and against any incitement to such discrimina
tion
Everyone has the right to an effective remedy by the competent national
tribunals for acts violating the fundamental rights granted him by the
constitution or by law

No one shall be subjected to arbitrary arrest detention or exile
Everyone is entitled in full equality to a fair and public hearing by a n independent and impartial tribunal in the determination of his rights and obligations and of any criminal charge against him
Everyone charged with a penal offence has the right to be presumed inno cent until proved guilty according to law in a public trial at which he has had all the guarantees necessary for his defence
No one shall be held guilty of any penal offence on account of any act or omission which did not constitute a penal offence under national or international law at the time when it was committed
Nor shall a heavier penalty be imposed than the one that was applicable at the time the penal offence was committed
No one shall be subjected to arbitrary interference with his privacy fa mily home or correspondence nor to attacks upon his honour and reputati on
Everyone has the right to the protection of the law against such interf erence or attacks
Everyone has the right to freedom of movement and residence within the borders of each State
Everyone has the right to leave any country including his own and to re turn to his country
Everyone has the right to seek and to enjoy in other countries asylum f rom persecution
This right may not be invoked in the case of prosecutions genuinely ari sing from non political crimes or from acts contrary to the purposes an d principles of the United Nations
Everyone has the right to a nationality
No one shall be arbitrarily deprived of his nationality nor denied the right to change his nationality
Men and women of full age without any limitation due to race nationalit y or religion have the right to marry and to found a family
They are entitled to equal rights as to marriage during marriage and at its dissolution
Marriage shall be entered into only with the free and full consent of t he intending spouses
The family is the natural and fundamental group unit of society and is entitled to protection by society and the State
Everyone has the right to own property alone as well as in association with others
No one shall be arbitrarily deprived of his property
Everyone has the right to freedom of thought conscience and religion
this right includes freedom to change his religion or belief and freedo m either alone or in community with others and in public or private to manifest his religion or belief in teaching practice worship and observ ance
Everyone has the right to freedom of opinion and expression
this right includes freedom to hold opinions without interference and t o seek receive and impart information and ideas through any media and r egardless of frontiers
Everyone has the right to freedom of peaceful assembly and association
No one may be compelled to belong to an association
Everyone has the right to take part in the government of his country di rectly or through freely chosen representatives
Everyone has the right to equal access to public service in his country
The will of the people shall be the basis of the authority of governmen t
this will shall be expressed in periodic and genuine elections which sh

all be by universal and equal suffrage and shall be held by secret vote or by equivalent free voting procedures
Everyone as a member of society has the right to social security and is entitled to realization through national effort and international co op eration and in accordance with the organization and resources of each S tate of the economic social and cultural rights indispensable for his d ignity and the free development of his personality
Everyone has the right to work to free choice of employment to just and favourable conditions of work and to protection against unemployment
Everyone without any discrimination has the right to equal pay for equa l work
Everyone who works has the right to just and favourable remuneration en suring for himself and his family an existence worthy of human dignity and supplemented if necessary by other means of social protection
Everyone has the right to form and to join trade unions for the protect ion of his interests
Everyone has the right to rest and leisure including reasonable limitat ion of working hours and periodic holidays with pay
Everyone has the right to a standard of living adequate for the health and well being of himself and of his family including food clothing hou sing and medical care and necessary social services and the right to se curity in the event of unemployment sickness disability widowhood old a ge or other lack of livelihood in circumstances beyond his control
Motherhood and childhood are entitled to special care and assistance
All children whether born in or out of wedlock shall enjoy the same soc ial protection
Everyone has the right to education
Education shall be free at least in the elementary and fundamental stag es
Elementary education shall be compulsory
Technical and professional education shall be made generally available and higher education shall be equally accessible to all on the basis of merit
Education shall be directed to the full development of the human person ality and to the strengthening of respect for human rights and fundamen tal freedoms
It shall promote understanding tolerance and friendship among all natio ns racial or religious groups and shall further the activities of the U nited Nations for the maintenance of peace
Parents have a prior right to choose the kind of education that shall b e given to their children
Everyone has the right freely to participate in the cultural life of th e community to enjoy the arts and to share in scientific advancement an d its benefits
Everyone has the right to the protection of the moral and material inte rests resulting from any scientific literary or artistic production of which he is the author
Everyone is entitled to a social and international order in which the r ights and freedoms set forth in this Declaration can be fully realized
Everyone has duties to the community in which alone the free and full d evelopment of his personality is possible
In the exercise of his rights and freedoms everyone shall be subject on ly to such limitations as are determined by law solely for the purpose of securing due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality public order an d the general welfare in a democratic society
These rights and freedoms may in no case be exercised contrary to the p

```
urposes and principles of the United Nations
Nothing in this Declaration may be interpreted as implying for any Stat
e group or person any right to engage in any activity or to perform any
act aimed at the destruction of any of the rights and freedoms set fort
h herein
```

In [5]:
```
docs = corpus.splitlines()    #Each document is one line of the corpus
```

In [6]:
```
print(docs)
```

['Whereas recognition of the inherent dignity and of the equal and inal
ienable rights of all members of the human family is the foundation of
freedom justice and peace in the world', 'Whereas disregard and contemp
t for human rights have resulted in barbarous acts which have outraged
the conscience of mankind and the advent of a world in which human bein
gs shall enjoy freedom of speech and belief and freedom from fear and w
ant has been proclaimed as the highest aspiration of the common peopl
e', 'Whereas it is essential if man is not to be compelled to have reco
urse as a last resort to rebellion against tyranny and oppression that
human rights should be protected by the rule of law', 'Whereas it is es
sential to promote the development of friendly relations between nation
s', 'Whereas the peoples of the United Nations have in the Charter reaf
firmed their faith in fundamental human rights in the dignity and worth
of the human person and in the equal rights of men and women and have d
etermined to promote social progress and better standards of life in la
rger freedom', 'Whereas Member States have pledged themselves to achiev
e in cooperation with the United Nations the promotion of universal res
pect for and observance of human rights and fundamental freedoms', 'Whe
reas a common understanding of these rights and freedoms is of the grea
test importance for the full realization of this pledge', 'Proclaims th
is Universal Declaration of Human Rights as a common standard of achiev
ement for all peoples and all nations to the end that every individual
and every organ of society keeping this Declaration constantly in mind
shall strive by teaching and education to promote respect for these rig
hts and freedoms and by progressive measures national and international
to secure their universal and effective recognition and observance both
among the peoples of Member States themselves and among the peoples of
territories under their jurisdiction', 'All human beings are born free
and equal in dignity and rights', 'They are endowed with reason and con
science and should act towards one another in a spirit of brotherhood',
'Everyone is entitled to all the rights and freedoms set forth in this
Declaration without distinction of any kind such as race colour sex lan
guage religion political or other opinion national or social origin pro
perty birth or other status', 'Furthermore no distinction shall be made
on the basis of the political jurisdictional or international status of
the country or territory to which a person belongs whether it be indepe
ndent trust non self governing or under any other limitation of soverei
gnty', 'Everyone has the right to life liberty and security of person',
'No one shall be held in slavery or servitude', 'slavery and the slave
trade shall be prohibited in all their forms', 'No one shall be subject
ed to torture or to cruel inhuman or degrading treatment or punishmen
t', 'Everyone has the right to recognition everywhere as a person befor
e the law', 'All are equal before the law and are entitled without any
discrimination to equal protection of the law', 'All are entitled to eq
ual protection against any discrimination in violation of this Declarat
ion and against any incitement to such discrimination', 'Everyone has t
he right to an effective remedy by the competent national tribunals for
acts violating the fundamental rights granted him by the constitution o
r by law', 'No one shall be subjected to arbitrary arrest detention or
exile', 'Everyone is entitled in full equality to a fair and public hea
ring by an independent and impartial tribunal in the determination of h
is rights and obligations and of any criminal charge against him', 'Eve
ryone charged with a penal offence has the right to be presumed innocen
t until proved guilty according to law in a public trial at which he ha
s had all the guarantees necessary for his defence', 'No one shall be h
eld guilty of any penal offence on account of any act or omission which
did not constitute a penal offence under national or international law

at the time when it was committed', 'Nor shall a heavier penalty be imp
osed than the one that was applicable at the time the penal offence was
committed', 'No one shall be subjected to arbitrary interference with h
is privacy family home or correspondence nor to attacks upon his honour
and reputation', 'Everyone has the right to the protection of the law a
gainst such interference or attacks', 'Everyone has the right to freedo
m of movement and residence within the borders of each State', 'Everyon
e has the right to leave any country including his own and to return to
his country', 'Everyone has the right to seek and to enjoy in other cou
ntries asylum from persecution', 'This right may not be invoked in the
case of prosecutions genuinely arising from non political crimes or fro
m acts contrary to the purposes and principles of the United Nations',
'Everyone has the right to a nationality', 'No one shall be arbitrarily
deprived of his nationality nor denied the right to change his national
ity', 'Men and women of full age without any limitation due to race nat
ionality or religion have the right to marry and to found a family', 'T
hey are entitled to equal rights as to marriage during marriage and at
its dissolution', 'Marriage shall be entered into only with the free an
d full consent of the intending spouses', 'The family is the natural an
d fundamental group unit of society and is entitled to protection by so
ciety and the State', 'Everyone has the right to own property alone as
well as in association with others', 'No one shall be arbitrarily depri
ved of his property', 'Everyone has the right to freedom of thought con
science and religion', 'this right includes freedom to change his relig
ion or belief and freedom either alone or in community with others and
in public or private to manifest his religion or belief in teaching pra
ctice worship and observance', 'Everyone has the right to freedom of op
inion and expression', 'this right includes freedom to hold opinions wi
thout interference and to seek receive and impart information and ideas
through any media and regardless of frontiers', 'Everyone has the right
to freedom of peaceful assembly and association', 'No one may be compel
led to belong to an association', 'Everyone has the right to take part
in the government of his country directly or through freely chosen repr
esentatives', 'Everyone has the right to equal access to public service
in his country', 'The will of the people shall be the basis of the auth
ority of government', 'this will shall be expressed in periodic and gen
uine elections which shall be by universal and equal suffrage and shall
be held by secret vote or by equivalent free voting procedures', 'Every
one as a member of society has the right to social security and is enti
tled to realization through national effort and international co operat
ion and in accordance with the organization and resources of each State
of the economic social and cultural rights indispensable for his dignit
y and the free development of his personality', 'Everyone has the right
to work to free choice of employment to just and favourable conditions
of work and to protection against unemployment', 'Everyone without any
discrimination has the right to equal pay for equal work', 'Everyone wh
o works has the right to just and favourable remuneration ensuring for
himself and his family an existence worthy of human dignity and supplem
ented if necessary by other means of social protection', 'Everyone has
the right to form and to join trade unions for the protection of his in
terests', 'Everyone has the right to rest and leisure including reasona
ble limitation of working hours and periodic holidays with pay', 'Every
one has the right to a standard of living adequate for the health and w
ell being of himself and of his family including food clothing housing
and medical care and necessary social services and the right to securit
y in the event of unemployment sickness disability widowhood old age or
other lack of livelihood in circumstances beyond his control', 'Motherh

```
ood and childhood are entitled to special care and assistance', 'All ch
ildren whether born in or out of wedlock shall enjoy the same social pr
otection', 'Everyone has the right to education', 'Education shall be f
ree at least in the elementary and fundamental stages', 'Elementary edu
cation shall be compulsory', 'Technical and professional education shal
l be made generally available and higher education shall be equally acc
essible to all on the basis of merit', 'Education shall be directed to
the full development of the human personality and to the strengthening
of respect for human rights and fundamental freedoms', 'It shall promot
e understanding tolerance and friendship among all nations racial or re
ligious groups and shall further the activities of the United Nations f
or the maintenance of peace', 'Parents have a prior right to choose the
kind of education that shall be given to their children', 'Everyone has
the right freely to participate in the cultural life of the community t
o enjoy the arts and to share in scientific advancement and its benefit
s', 'Everyone has the right to the protection of the moral and material
interests resulting from any scientific literary or artistic production
of which he is the author', 'Everyone is entitled to a social and inter
national order in which the rights and freedoms set forth in this Decla
ration can be fully realized', 'Everyone has duties to the community in
which alone the free and full development of his personality is possibl
e', 'In the exercise of his rights and freedoms everyone shall be subje
ct only to such limitations as are determined by law solely for the pur
pose of securing due recognition and respect for the rights and freedom
s of others and of meeting the just requirements of morality public ord
er and the general welfare in a democratic society', 'These rights and
freedoms may in no case be exercised contrary to the purposes and princ
iples of the United Nations', 'Nothing in this Declaration may be inter
preted as implying for any State group or person any right to engage in
any activity or to perform any act aimed at the destruction of any of t
he rights and freedoms set forth herein']
```

Tokenization

In [7]:
```
doc_tokens = [x.split() for x in docs]  #Defining tokens as words

nltk.download('punkt')
doc_tokens_2 = [nltk.word_tokenize(x) for x in docs]  #Subtle difference
s, particularly: "in, a" --> ["in,", "a"]
                                                      #as opposed to
 "in, a" --> ["in", ",", "a"]
```

```
[nltk_data] Downloading package punkt to /Users/cgb/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

In [8]:
```
doc_tokens[0]
```

Out[8]:
```
['Whereas',
 'recognition',
 'of',
 'the',
 'inherent',
 'dignity',
 'and',
 'of',
 'the',
 'equal',
 'and',
 'inalienable',
 'rights',
 'of',
 'all',
 'members',
 'of',
 'the',
 'human',
 'family',
 'is',
 'the',
 'foundation',
 'of',
 'freedom',
 'justice',
 'and',
 'peace',
 'in',
 'the',
 'world']
```

Stopword removal

In [9]:
```
nltk.download('stopwords')
from nltk.corpus import stopwords
stop = stopwords.words('english')
```

```
[nltk_data] Downloading package stopwords to /Users/cgb/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [10]:
```
doc_tokens_clean = [[x.lower() for x in words if x.lower() not in stop]
for words in doc_tokens]
   #Make all tokens lowercase and filter out stopwords
```

In [11]:
```
doc_tokens_clean[0]
```

Out[11]:
```
['whereas',
 'recognition',
 'inherent',
 'dignity',
 'equal',
 'inalienable',
 'rights',
 'members',
 'human',
 'family',
 'foundation',
 'freedom',
 'justice',
 'peace',
 'world']
```

Lemmatizing

In [13]:
```
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer

stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

doc_tokens_clean_lem = [[lemmatizer.lemmatize(x) for x in words] for wor
ds in doc_tokens_clean]
```

In [14]:
```
doc_tokens_clean[1]
```

Out[14]:
```
['whereas',
 'disregard',
 'contempt',
 'human',
 'rights',
 'resulted',
 'barbarous',
 'acts',
 'outraged',
 'conscience',
 'mankind',
 'advent',
 'world',
 'human',
 'beings',
 'shall',
 'enjoy',
 'freedom',
 'speech',
 'belief',
 'freedom',
 'fear',
 'want',
 'proclaimed',
 'highest',
 'aspiration',
 'common',
 'people']
```

In [15]:
```
doc_tokens_clean_lem[1]
```

Out[15]:
```
['whereas',
 'disregard',
 'contempt',
 'human',
 'right',
 'resulted',
 'barbarous',
 'act',
 'outraged',
 'conscience',
 'mankind',
 'advent',
 'world',
 'human',
 'being',
 'shall',
 'enjoy',
 'freedom',
 'speech',
 'belief',
 'freedom',
 'fear',
 'want',
 'proclaimed',
 'highest',
 'aspiration',
 'common',
 'people']
```

Stemming vs. Lemmatizing

In [16]:
```
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

#The lemmatizer will assume we want the word lemmatized to a noun unless we specify the part of speech (POS)
#Changing the POS tag will then change the result we get
def show_words(words):
    for w, pos in words:
        print(f'Word: {w:10}, Stem: {stemmer.stem(w):10}, Lemma: {lemmatizer.lemmatize(w, pos):10}')
show_words([('stones', 'n'), ('jokes', 'n')])
```

```
Word: stones    , Stem: stone     , Lemma: stone
Word: jokes     , Stem: joke      , Lemma: joke
```

In [17]:
```
show_words([('speak', 'v'), ('speaking', 'v'), ('spoken', 'v')])
```

```
Word: speak     , Stem: speak     , Lemma: speak
Word: speaking  , Stem: speak     , Lemma: speak
Word: spoken    , Stem: spoken    , Lemma: speak
```

In [18]:
```
show_words([('spoke', 'v'), ('spoke', 'n')])
```

```
Word: spoke      , Stem: spoke      , Lemma: speak
Word: spoke      , Stem: spoke      , Lemma: spoke
```

In [19]:
```
show_words([('foot', 'n'), ('feet', 'n'),  ('goose', 'n'), ('geese', 'n'
)])
```

```
Word: foot       , Stem: foot       , Lemma: foot
Word: feet       , Stem: feet       , Lemma: foot
Word: goose      , Stem: goos       , Lemma: goose
Word: geese      , Stem: gees       , Lemma: goose
```

In [20]:
```
show_words([('is', 'v'), ('are', 'v'), ('be', 'v')])
```

```
Word: is         , Stem: is         , Lemma: be
Word: are        , Stem: are        , Lemma: be
Word: be         , Stem: be         , Lemma: be
```

Document-word matrix

In [21]:
```
#A simple way of building a document-word matrix
word_list = []
for doc in doc_tokens_clean_lem:
    for word in doc:
        if(not(word in word_list)):
            word_list.append(word)
doc_word_simple = []
for doc in doc_tokens_clean_lem:
    doc_vec = [0]*len(word_list)  #Each document is represented as a vec
tor of word occurrences
    for word in doc:
        ind = word_list.index(word)
        doc_vec[ind] += 1    #Increment the corresponding word index
    doc_word_simple.append(doc_vec)
```

In [22]:
```
doc_word_simple[0][:10]
```

Out[22]:
```
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

In [23]:
```
doc_word_simple[2][:10]
```

Out[23]:
```
[1, 0, 0, 0, 0, 0, 1, 0, 1, 0]
```

In [24]:
```
doc_word_simple = np.array(doc_word_simple)   #Now we can use numpy oper
ations on the matrix
```

In [25]:
```
doc_word_simple
```

Out[25]:
```
array([[1, 1, 1, ..., 0, 0, 0],
       [1, 0, 0, ..., 0, 0, 0],
       [1, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 1, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 1, 1, 1]])
```

In [26]:
```
# Faster with some optimizations
# Create dictionary so faster lookup
# Allocate memory ahead of time via numpy
word_to_ind = {word:ind for ind, word in enumerate(word_list)}
doc_word = np.zeros((len(doc_tokens_clean_lem), len(word_list)))
for doc, doc_vec in zip(doc_tokens_clean_lem, doc_word):
    for word in doc:
        ind = word_to_ind[word]
        doc_vec[ind] += 1

# Check that this produces the same result
np.all(np.isclose(doc_word, doc_word_simple))
```

Out[26]:
```
True
```

In [ ]:

In [ ]:

In [ ]: